# Groupwork2

Kuan

2023-03-10

```
library(readr)
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(janitor)
library(areaplot)
library(dplyr)
library(skimr)
library(kableExtra)
library(gridExtra)
library(ggplot2)
library(stringr)
```

## Data pre-processing

### Remove missing value

In the raw dataset, there are some missing data about mean altitude and harvested, so before analysis data we remove missing values.

```
dataset13 <- read_csv("dataset13.csv")
dim(dataset13)
```

```
[1] 1145    8
```

```
#remove NA
newdataset<- na.omit(dataset13)
dim(newdataset)
```

```
[1] 935   8
```

### data cleaning

The boxplot about mean altitude shows that there are some outliers. Four of them are more than 10000 metres, obviously they are wrong datas, so we remove them.From the boxplot about Aroma, we find there is a wrong value which equals zero and remove it from the dataset.

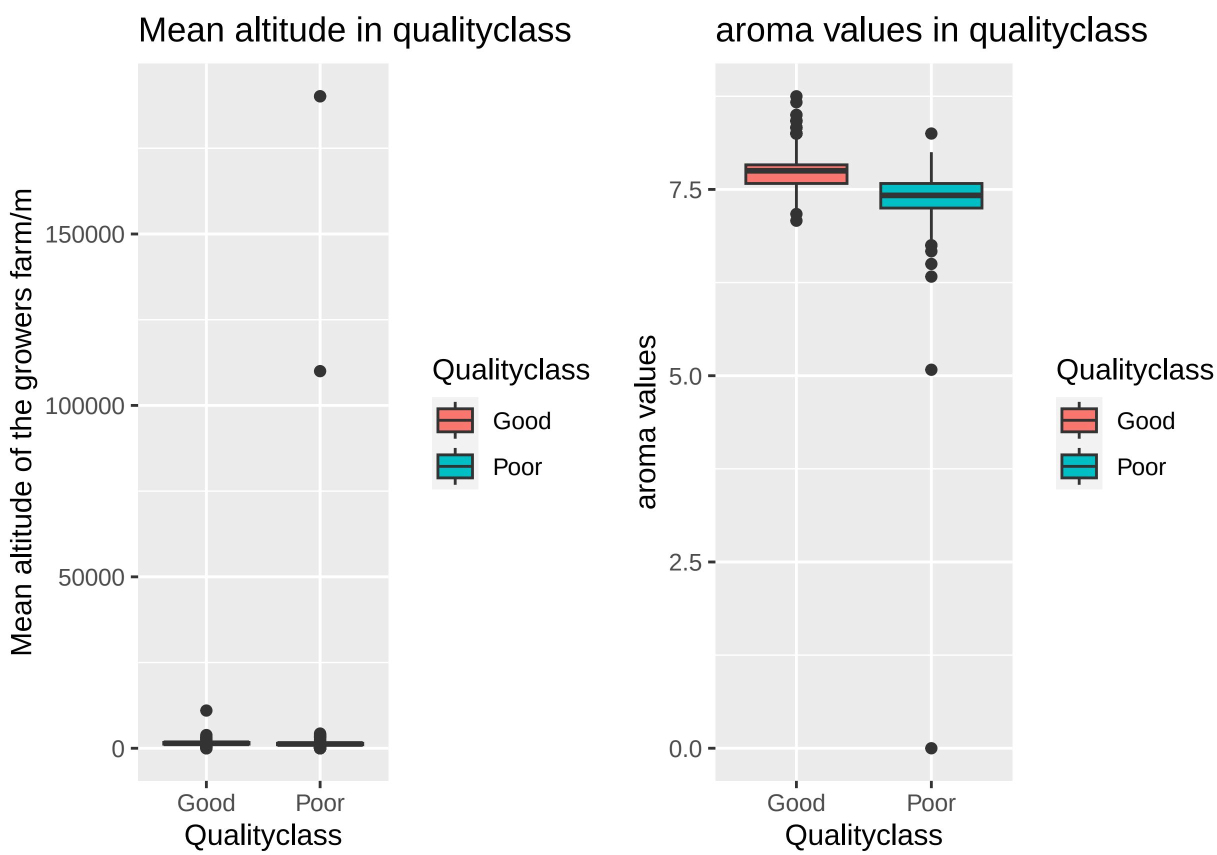


Figure 1: Boxplots of mean altitude of the growers farm/left, aroma values/right in different qualityclass

## Suitable numerical summaries and visualizations

```
#summary of numerical explanatory variables
newdata_summary<- newdataset%>%
  dplyr::select(aroma,flavor,acidity,category_two_defects,altitude_mean_meters)
my_skim <- skim_with(numeric = sfl(hist = NULL),
                    base = sfl(n = length))
my_skim(newdata_summary) %>%
  transmute(Variable=skim_variable, n = n, Mean=numeric.mean, SD=numeric.sd,
            Min=numeric.p0, Median=numeric.p50,  Max=numeric.p100,
            IQR = numeric.p75-numeric.p50) %>%
  kable(caption = '\\label{tab:summaries1} Summary statistics on the different numerical explanatory va
  kable_styling(font_size = 10, latex_options = "HOLD_position")
```

Table 1: Summary statistics on the different numerical explanatory variables of coffee.

| Variable | n | Mean | SD | Min | Median | Max | IQR |
|---|---|---|---|---|---|---|---|
| aroma | 930 | 7.58 | 0.31 | 5.08 | 7.58 | 8.75 | 0.17 |
| flavor | 930 | 7.53 | 0.32 | 6.17 | 7.58 | 8.67 | 0.17 |
| acidity | 930 | 7.53 | 0.31 | 5.25 | 7.50 | 8.58 | 0.25 |
| category_two_defects | 930 | 3.64 | 5.35 | 0.00 | 2.00 | 47.00 | 2.00 |
| altitude_mean_meters | 930 | 1325.65 | 484.31 | 1.00 | 1310.64 | 4287.00 | 289.36 |

Table1 shows that the mean values of Aroma grade, Flavor grade and Acidity grade are both approximately 7.5. There are large differences of category two defects between different coffee beans, as some have no defective product, but some have 47 in the batch of coffee beans tested. Similarly, the difference in mean altitude is distinct.

```
#summary of categorical explanatory variables
#country of origin
data_country<- newdataset %>%
  group_by(country_of_origin) %>%
  summarise(n=n())
data_country
```

```
# A tibble: 33 x 2
   country_of_origin     n
   <chr>             <int>
 1 Brazil               91
 2 Burundi               2
 3 China                14
 4 Colombia            127
 5 Costa Rica           36
 6 Cote d?Ivoire         1
 7 Ecuador               2
 8 El Salvador          18
 9 Ethiopia             23
10 Guatemala           127
# ... with 23 more rows
```

```r
newdataset %>%
  tabyl(country_of_origin, Qualityclass) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns() %>%
  kable(caption = '\\label{tab1:origin} Summary statistics on country of origin.') %>%
  kable_styling(latex_options = "HOLD_position")
```

Table 2: Summary statistics on country of origin.

| country_of_origin | Good | Poor |
|---|---|---|
| Brazil | 51.6% (47) | 48.4% (44) |
| Burundi | 50.0% (1) | 50.0% (1) |
| China | 64.3% (9) | 35.7% (5) |
| Colombia | 82.7% (105) | 17.3% (22) |
| Costa Rica | 55.6% (20) | 44.4% (16) |
| Cote d?Ivoire | 0.0% (0) | 100.0% (1) |
| Ecuador | 50.0% (1) | 50.0% (1) |
| El Salvador | 72.2% (13) | 27.8% (5) |
| Ethiopia | 100.0% (23) | 0.0% (0) |
| Guatemala | 52.0% (66) | 48.0% (61) |
| Haiti | 20.0% (1) | 80.0% (4) |
| Honduras | 26.1% (12) | 73.9% (34) |
| India | 50.0% (5) | 50.0% (5) |
| Indonesia | 57.1% (8) | 42.9% (6) |
| Kenya | 90.0% (18) | 10.0% (2) |
| Laos | 0.0% (0) | 100.0% (2) |
| Malawi | 9.1% (1) | 90.9% (10) |
| Mauritius | 0.0% (0) | 100.0% (1) |
| Mexico | 26.0% (52) | 74.0% (148) |
| Myanmar | 0.0% (0) | 100.0% (6) |
| Nicaragua | 23.1% (3) | 76.9% (10) |
| Panama | 75.0% (3) | 25.0% (1) |
| Peru | 0.0% (0) | 100.0% (1) |
| Philippines | 40.0% (2) | 60.0% (3) |
| Taiwan | 40.4% (23) | 59.6% (34) |
| Tanzania, United Republic Of | 48.3% (14) | 51.7% (15) |
| Thailand | 57.1% (8) | 42.9% (6) |
| Uganda | 76.7% (23) | 23.3% (7) |
| United States | 66.7% (6) | 33.3% (3) |
| United States (Hawaii) | 100.0% (1) | 0.0% (0) |
| United States (Puerto Rico) | 33.3% (1) | 66.7% (2) |
| Vietnam | 57.1% (4) | 42.9% (3) |
| Zambia | 0.0% (0) | 100.0% (1) |

The summary table shows that there are total 33 countries in the dataset, and 200 observations are from Mexico, which is the most, but some countries have only one observation. We also note that there are 6 countries like Laos only have poor qualityclass of coffee, the qualityclass of Ethiopia and United States(Hawaii) are all good. There also have 3 countries' qualityclass is half and half.

```r
#harvested
data_harvested<- newdataset %>%
  group_by(harvested) %>%
  summarise(n=n())
data_harvested
```

```
# A tibble: 9 x 2
  harvested     n
      <dbl> <int>
1      2010    26
2      2011    30
3      2012   255
4      2013   134
5      2014   194
6      2015   118
7      2016   103
8      2017    52
9      2018    18
```

```r
newdataset %>%
  tabyl(harvested, Qualityclass) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns() %>%
  kable(caption = '\\label{tab1:harvested} Summary statistics on
harvested.') %>%
  kable_styling(latex_options = "HOLD_position")
```

Table 3: Summary statistics on harvested.

| harvested | Good | Poor |
|---|---|---|
| 2010 | 76.9% (20) | 23.1% (6) |
| 2011 | 73.3% (22) | 26.7% (8) |
| 2012 | 39.2% (100) | 60.8% (155) |
| 2013 | 53.7% (72) | 46.3% (62) |
| 2014 | 50.0% (97) | 50.0% (97) |
| 2015 | 54.2% (64) | 45.8% (54) |
| 2016 | 55.3% (57) | 44.7% (46) |
| 2017 | 48.1% (25) | 51.9% (27) |
| 2018 | 72.2% (13) | 27.8% (5) |

The summary table shows that the information is collected from 2010 to 2018, and 255 observations is from 2012 which is the most. We also note that in 2010 the propotion of good qualityclass is highest, which is 76.9%. The lowest is 39.2% in 2012.

## Country of origin

```r
ggplot(newdataset, aes(x=Qualityclass, y=..prop.., group=country_of_origin, fill=country_of_origin))+
  geom_bar(position = "dodge", stat="count")+
  labs(y="Proportion")
```

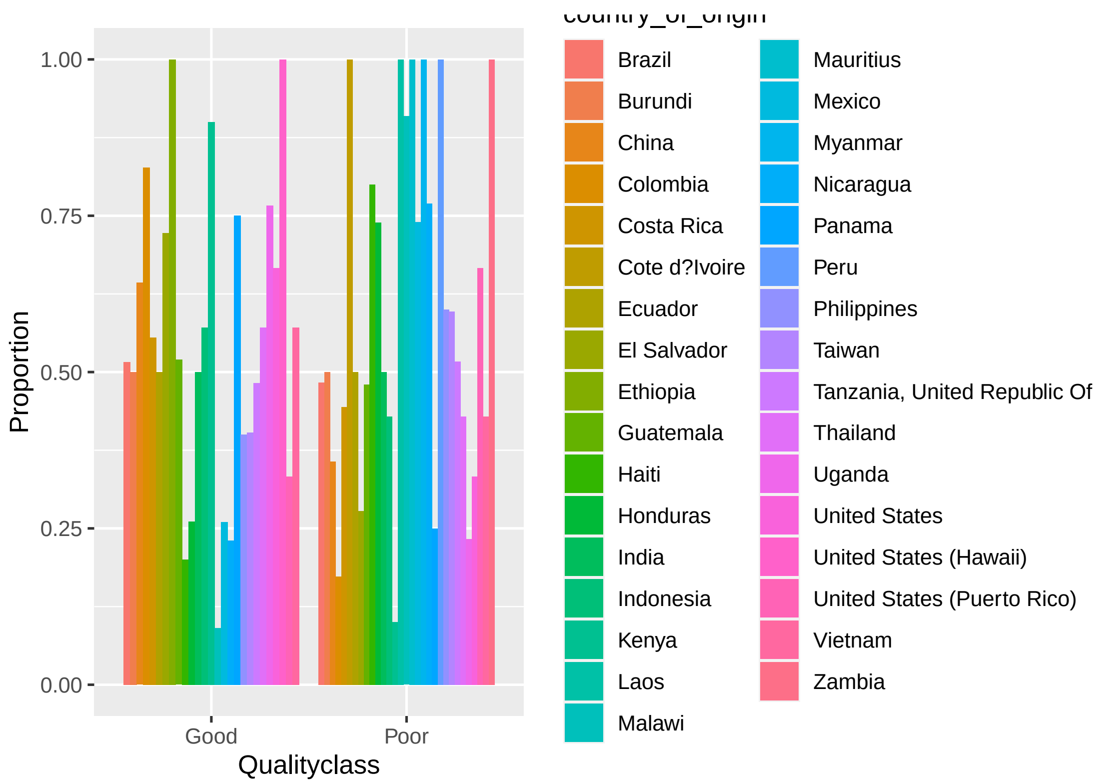


Figure 2: Propotion of Qualityclass by countries of origin.

The plot shows the propotion of Good qualityclass vs poor qualityclass between different country of origin. we can see some countries are all good quality, some are all poor quality.So we can fit a logistic regression model to determine whether the qualityclass of coffee can be predicted from their country of origin.

### Aroma, Flavour and Acidity

```r
#boxplot of Aroma grade
ggplot(data = newdataset, mapping = aes(x = factor(Qualityclass), y = aroma, fill = Qualityclass)) +
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Aroma",
       title = "Aroma in different qualityclass")
```

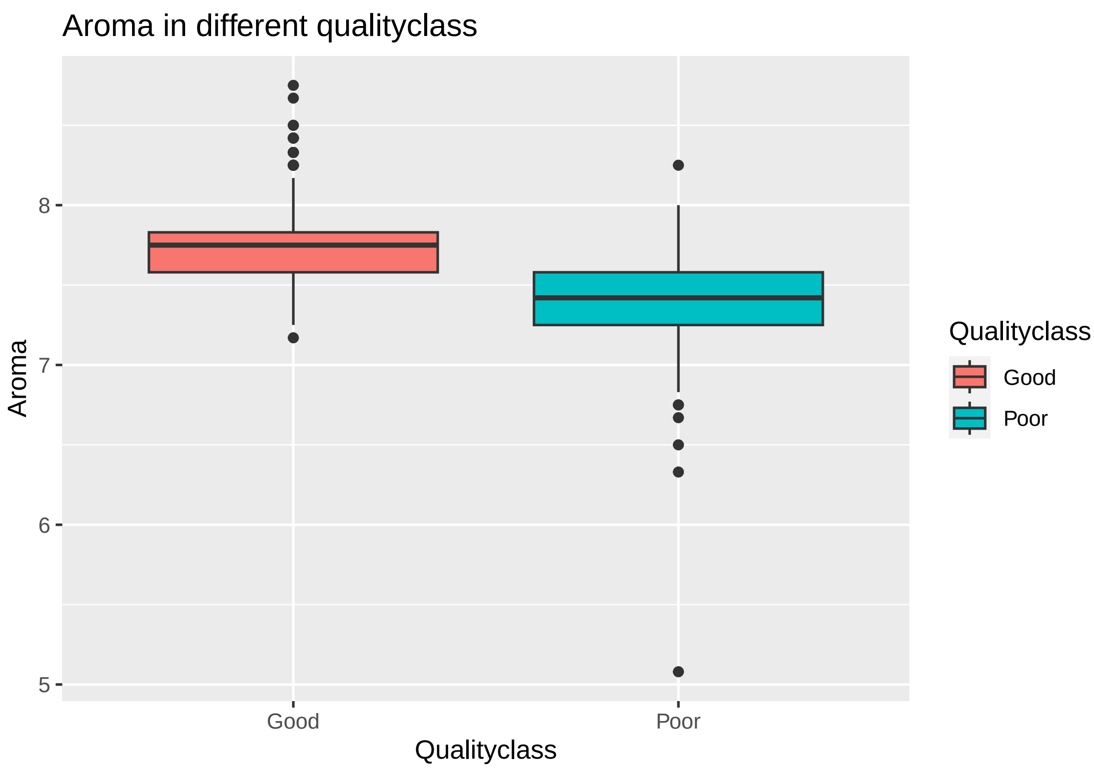


Figure 3: Boxplot of aroma in different qualityclass

```
#boxplot of Flavour grade
ggplot(data = newdataset, mapping = aes(x = factor(Qualityclass), y = flavor, fill = Qualityclass)) +
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Flavor",
       title = "Flavor in different qualityclass")
```

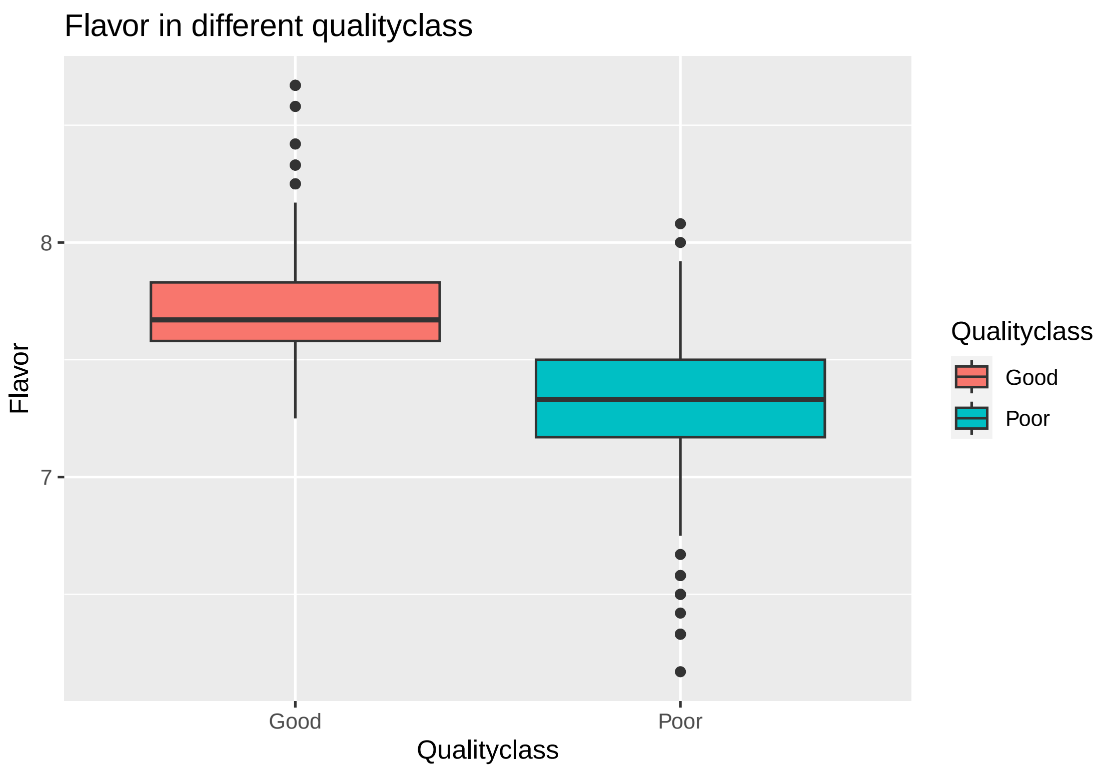


Figure 4: Boxplot of aroma in different qualityclass

```
#boxplot of Acidity grade
ggplot(data = newdataset, mapping = aes(x = factor(Qualityclass), y = acidity, fill = Qualityclass)) +
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Acidity",
       title = "Acidity in different qualityclass")
```

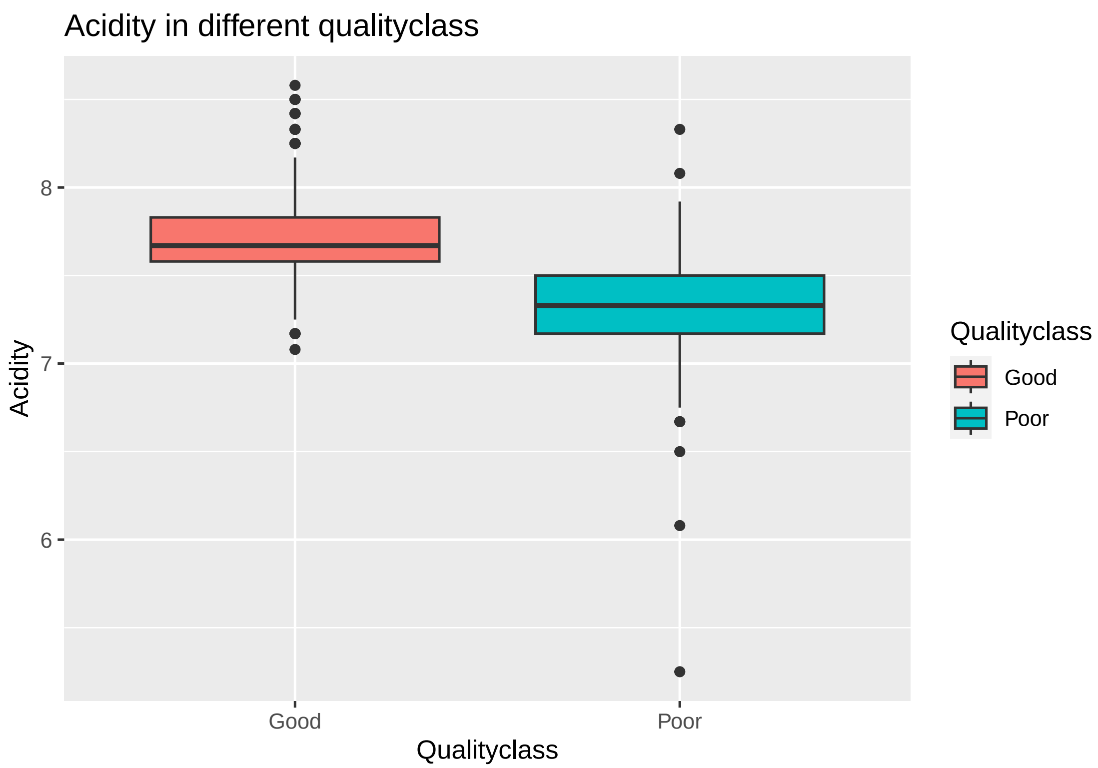


Figure 5: Boxplot of aroma in different qualityclass

The features of these three quality scores are similar. The boxplots show that coffee with good quality have higher grade in Aroma, Flavour and Acidity than poor. So we can fit a logistic regression model to see whether Aroma, Flavour and Acidity are significant predictors of the odds of qualityclass of coffee beans.

### Count of defects

```
ggplot(data = newdataset, mapping = aes(x = factor(Qualityclass), y = category_two_defects, fill = Quali
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Count of category 2 type defects",
       title = "Count of category 2 type defects in different qualityclass")
```

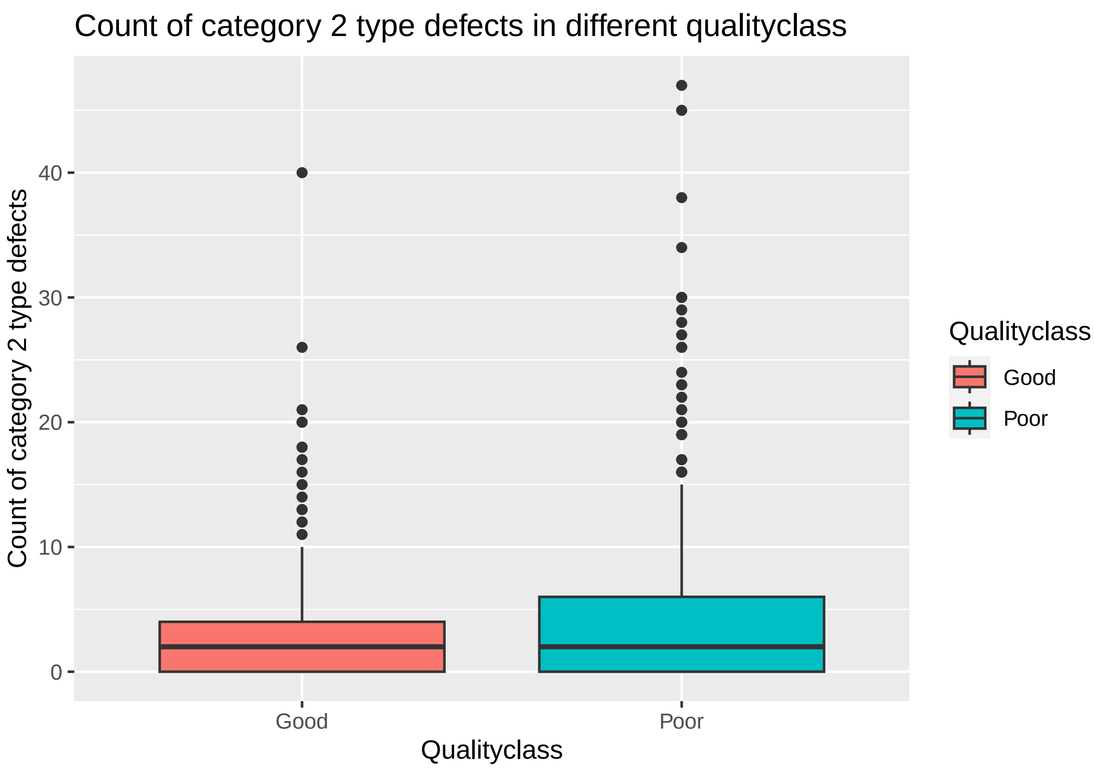


Figure 6: Boxplot of Count of category 2 type defects in different qualityclass

The boxplot shows that the poor quality coffee beans have more defective products than good quality ones, and there are more outliers in poor quality coffee beans.So we can fit a logistic regression model to see whether count of category 2 type defects is a significant predictor of the odds of qualityclass of coffee beans.

### Mean altitude

```
ggplot(data = newdataset, mapping = aes(x = factor(Qualityclass), y = altitude_mean_meters, fill = Quali
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Mean altitude of the growers farm/m",
       title = "Mean altitude of the growers farm in different qualityclass")
```

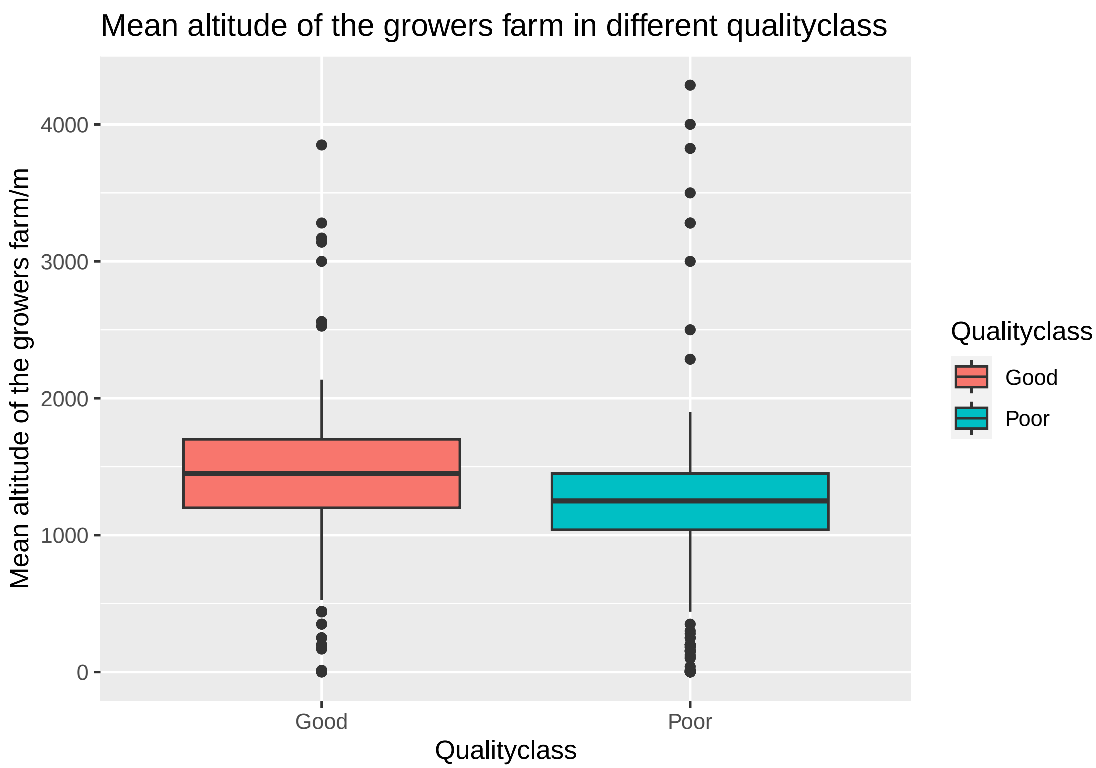


Figure 7: Boxplot of mean altitude of the growers farm in different qualityclass

The boxplot shows that the mean altitude of good quality coffee beans are higher than poor quality ones. we can notice that the poor quality have more outliers. So we can fit a logistic regression model to see whether mean altitude of the growers is a significant predictor of the odds of qualityclass of coffee beans.

### harvested

```
ggplot(newdataset, aes(x=Qualityclass, y=..prop.., group=harvested, fill=harvested))+
  geom_bar(position = "dodge", stat="count")+
  labs(y="Proportion")
```

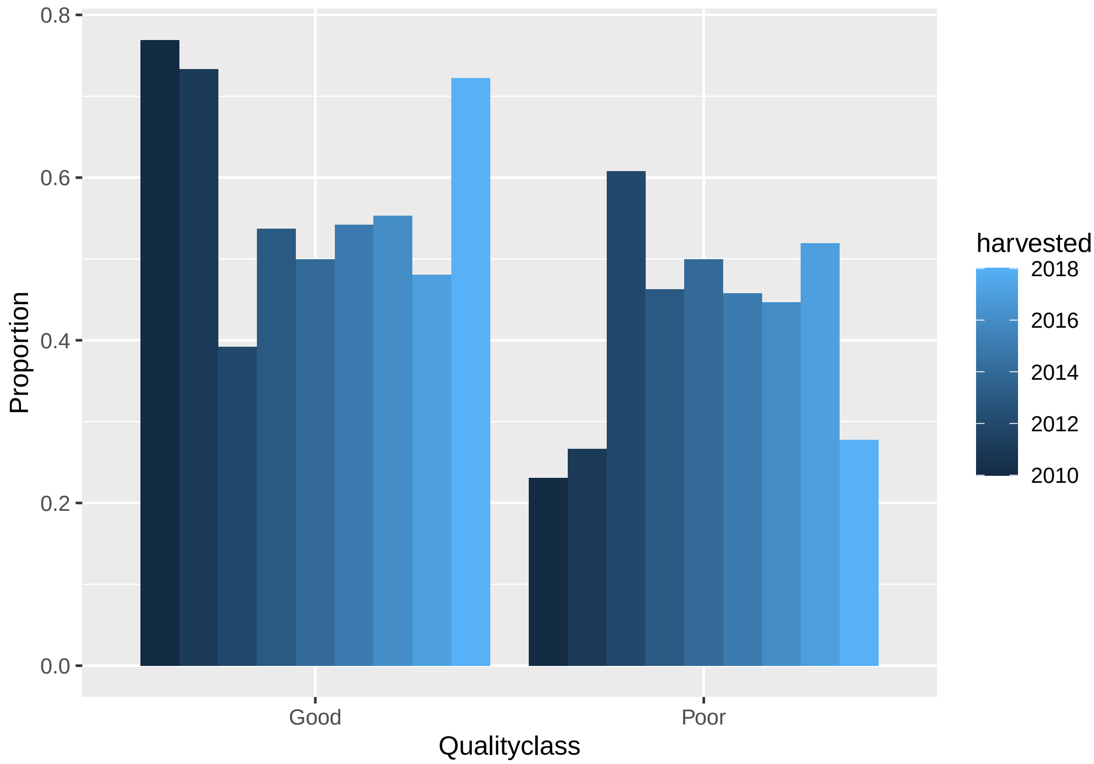


Figure 8: Propotion of Qualityclass by Harvested.

```
prop<-newdataset %>%
  tabyl(harvested, Qualityclass) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns()
prop$Good<-str_sub(string=prop$Good, start=1, end=5)
prop$Good<-as.factor(prop$Good)
ggplot(prop, aes(x=harvested, y=Good, group=1))+
  geom_line()+
  labs(y="Proportion")
```

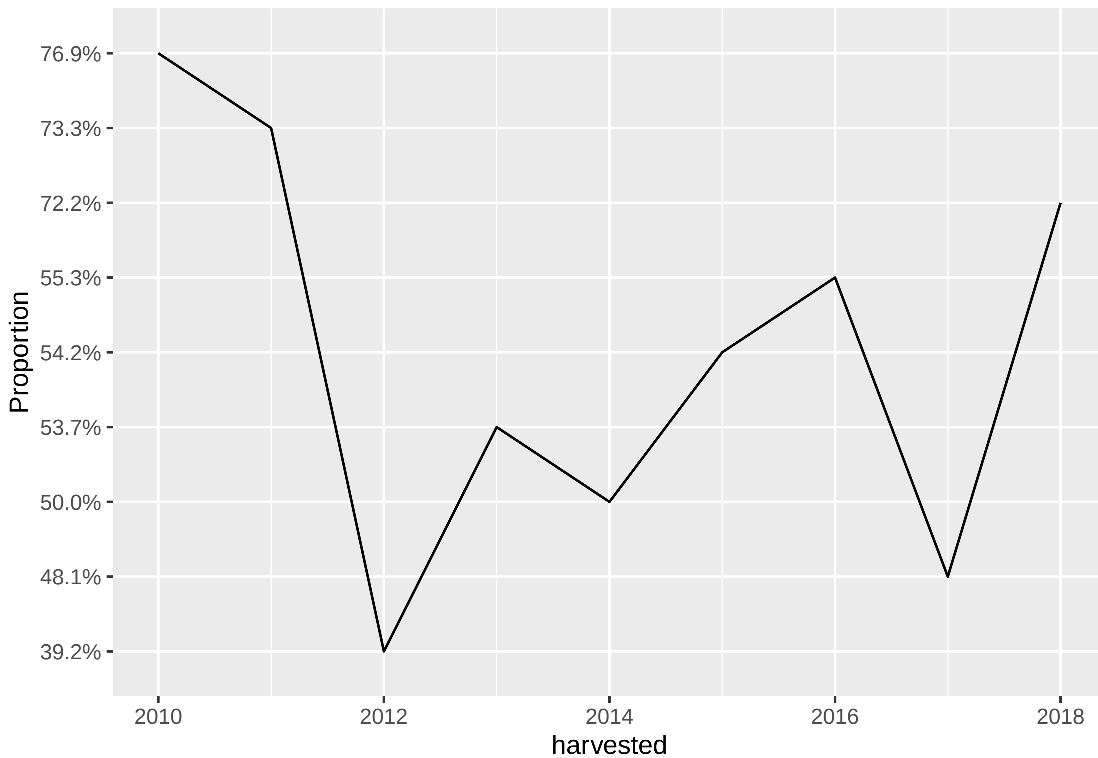


Figure 9: Propotion of Qualityclass by Harvested.

The line plot shows the propotion of Good qualityclass is highest in 2010,which is about 75%. We can fit a logistic regression model to determine whether the qualityclass of coffee can be predicted from harvested years.